Feature extraction from exchange data

Bogdan Alexandrov, Nikolay Ivanov
Skolkovo Institute of Science and Technology, Skolkovo, Russia

Skoltech
Skolkovo Institute of Science and Technology

High Performance Python Lab, 2023

20.12.2023

Skoltech
Skolkovo Institute of Science and Technology

Outline

Skoltech
Skolkovo Institute of Science and Technology

Motivation

Now days exchanges all over the world produces a lot of data, since there exists large amount of traders, especially on liquidity markets. Also, HFT traders, as it leads from name, work with the data of really high frequency - they retrieve order book snapshots and trades every microsecond. Hence, only one day of exchange dynamics can produce several gigabytes of data. And for evaluating strategies during backtest, it is needed to calculate features based on trades and best limit orders information. If this procedure will take a lot of time, traders won't be able to work efficiently and backtest algorithm will be quite time consuming, which will lead to fund money losing.

Skoltech
Skolkovo Institute of Science and Technology

Dataset

- Trades, ticker, order book raw data
- 319851, 1144985, 177479 rows of data
- Ticker contains best asks and bids prices and their amount
- Trades contains price of exchange and amount
- Order book contains bids and asks price and amount for best 25 price levels

Skoltech
Skolkovo Institute of Science and Technology

We want to get trading features

- Improved order book imbalances
- Autocorrelation
- Realized kernel
- and some other...

Skoltech
Skolkovo Institute of Science and Technology

Order Book Imbalances Problem

The order book imbalances problem refers to the situation in financial markets where there is a significant disparity between the buy and sell orders at a particular price level. In a typical order book, buy and sell orders are displayed, and the order book imbalance is calculated as the difference between the quantities of buy and sell orders. The classical notion of orderbook imbalance is given by:

$$\frac{b-a}{b+a} \in [-1, 1] \tag{1}$$

where a and b represent the top of the ask and bid liquidity, respectively

Skoltech
Skolkovo Institute of Science and Technology

Improved OrderBook Imbalanced Feature

Let us define the following quantities for each market $i = 1, \ldots, 14$ and for fixed quantities $N_1, \ldots, N_{14} \in R_1$

$$IMB_t^{a,i} = \left(\frac{p_{a,t}^i(N_i)}{p_{a,t}^i(1)} - 1\right) * 1000 \tag{2}$$

$$IMB_t^{b,i} = \left(\frac{p_{b,t}^i(1)}{p_{b,t}^i(N_i)} - 1\right) * 1000 \tag{3}$$

where $p_{a,t}^i(x)$, denotes the average price one would pay at time t for a market buy order of size x on market i

Skoltech
Skolkovo Institute of Science and Technology

Past return

Following a similar logic as above for the trade flow imbalance, it is to be expected that a drastic price change on one market should cross-impact short-term future returns on another market. It is therefore natural to include past returns on each market in our feature set. We use the same time horizons here as for the trade flow imbalances. Formally, for each market $i = 1, ...14$ and time horizon $\in$ 100ms, 250ms, 500ms, 1000ms, 2000ms, we define the following indicator at each time t

$$PRET_t^{a,\delta} = (\frac{p_t^i}{p_{t-\delta}^i} - 1) * 1000 \tag{4}$$

where p_t^i denotes the price on market i at time t

Skoltech
Skolkovo Institute of Science and Technology

$$p_t^i := \frac{1}{A} \sum_{(a,p) \in T_t^i} a \cdot p, \quad (11) \tag{5}$$

where T_t^i is the set of all trades, i.e., pairs (a, p) of amount and price, executed on market i in the time period $[t - 50\text{ms}, t]$, and $A = \sum_{(a,p) \in T_t^i} a$. Note that we include both buy and sell trades in T_t^i.

Skoltech
Skolkovo Institute of Science and Technology

Autocorrelation

autocorrelation function (ACF) at lag k for a time series (z_t)

$$-ac_k = \frac{E[(z_t - \mu)(z_{t+k} - \mu)]}{\sqrt{E[(z_t - \mu)^2]E[(z_{t+k} - \mu)^2]}} \tag{6}$$

Where ac_k represents the autocorrelation at lag k, z_t is the value of time series at time t, μ is mean of time serires, z_{t+k} value of time series at a future time $t + k$, where k is lag The autocorrelation at lag k is essentially a normalized measure of how correlated the values of the time series are at time t and $t + k$, considering the deviations from the mean

Skoltech
Skolkovo Institute of Science and Technology

Realized Kernel

Realized kernels are used to obtain a noise robust estimate of QV as follows:

$$RK_t = \gamma_0(X_n) + \sum_{i=1}^{H} k(\frac{h}{H})(\gamma_h(X_{\Delta_n}) + \gamma_{-h}(X_{\Delta_n})) \tag{7}$$

with H the kernel bandwidth, $\gamma_h(X_{\Delta_n})$ the autocovariation process, k is the kernel function of choice.

Skoltech
Skolkovo Institute of Science and Technology

Used Approaches

- CuPy
- Numba
- Numpy
- MPI (not for all)

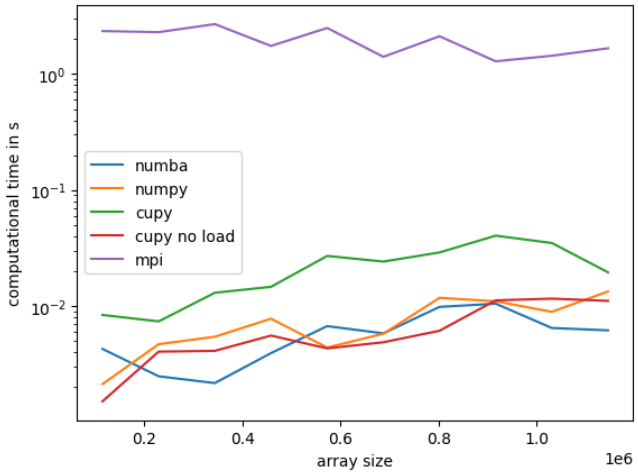
Skoltech
Skolkovo Institute of Science and Technology

Numba/CuPy/MPI

- Numba is a Just-In-Time (JIT) compiler for Python that translates functions into machine code at runtime.
- CuPy is a GPU-accelerated library for numerical computations in Python.
- It is built on top of CUDA, enabling seamless GPU acceleration with a NumPy-like interface.
- MPI (Message Passing Interface)
- Standard for parallel programming
- Enables communication between multiple processes

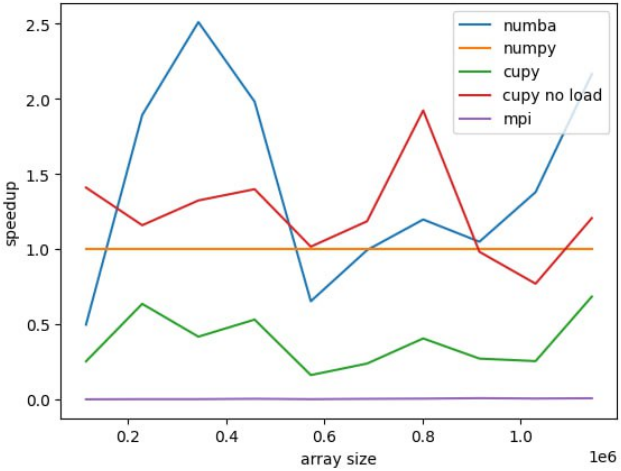
Skoltech
Skolkovo Institute of Science and Technology

Order book imbalance



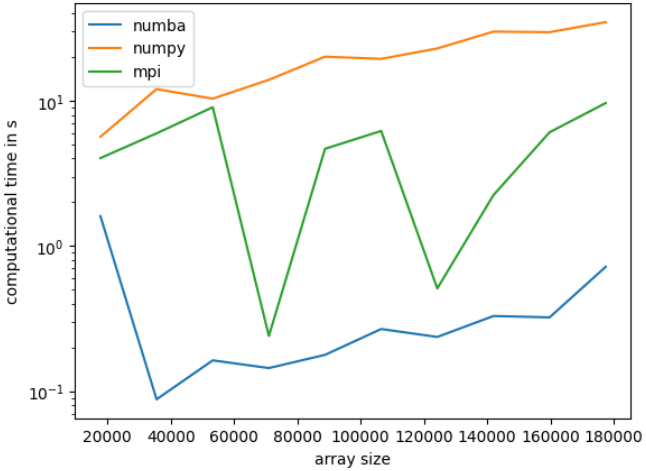
Skoltech
Skolkovo Institute of Science and Technology

Order book imbalance



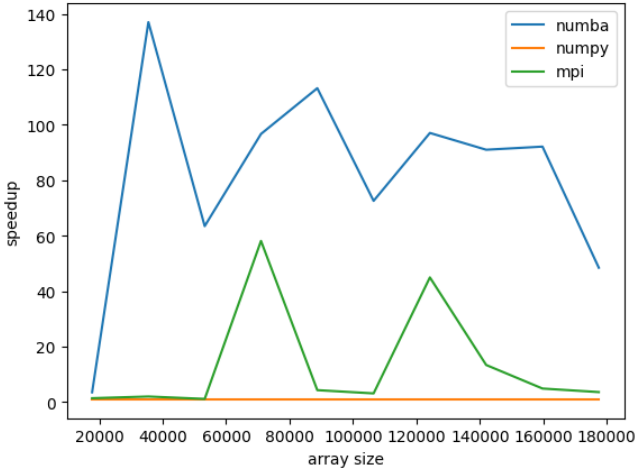
Skoltech
Skolkovo Institute of Science and Technology

Improved book imbalance



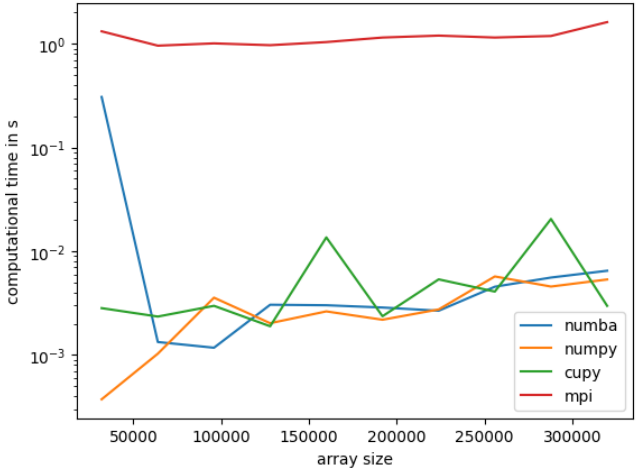
Skoltech
Skolkovo Institute of Science and Technology

Improved book imbalance



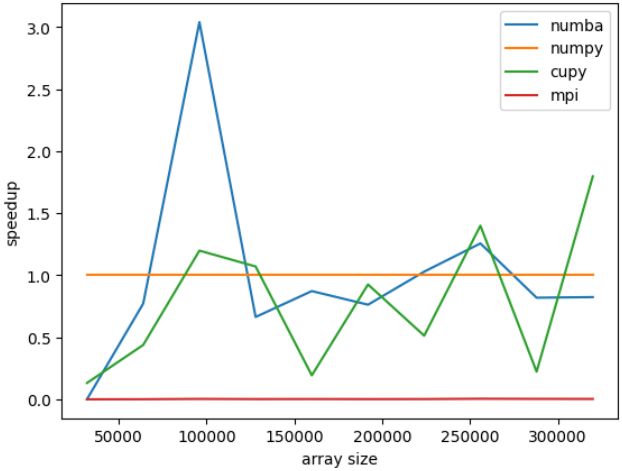
Skoltech
Skolkovo Institute of Science and Technology

Log Return



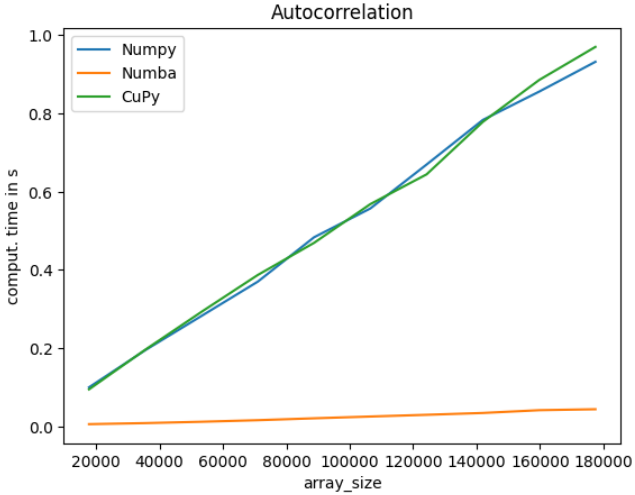
Skoltech
Skolkovo Institute of Science and Technology

Log Return



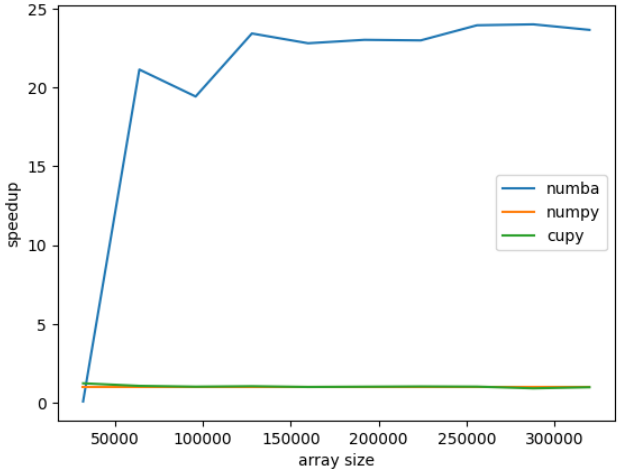
Skoltech
Skolkovo Institute of Science and Technology

Autocorrelation



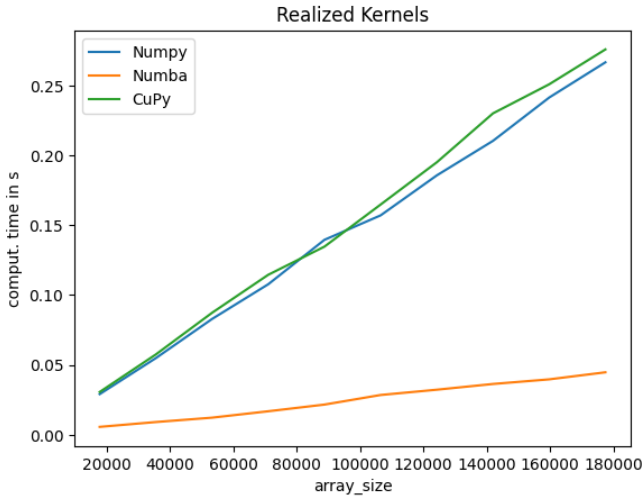
Skoltech
Skolkovo Institute of Science and Technology

Autocorrelation



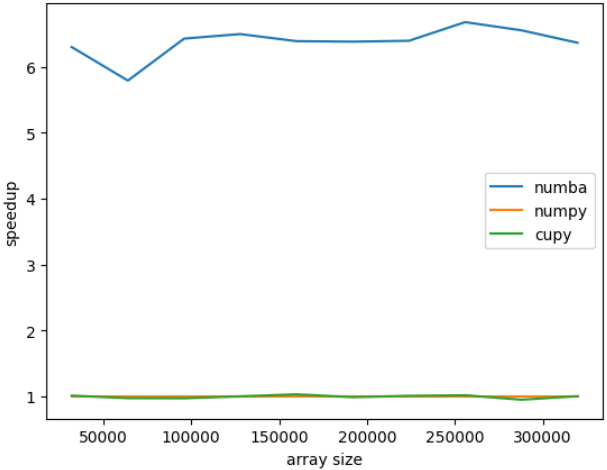
Skoltech
Skolkovo Institute of Science and Technology

Realized Kernel



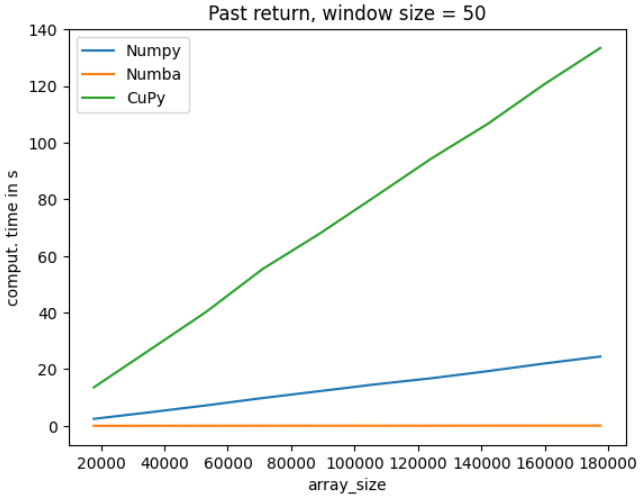
Skoltech
Skolkovo Institute of Science and Technology

Realized Kernel



Skoltech
Skolkovo Institute of Science and Technology

Past Return



Skoltech
Skolkovo Institute of Science and Technology

Conclusions & Future work

- Numba work perfrectly especially after compilations
- MPI is not applicable for all features

Skoltech
Skolkovo Institute of Science and Technology

Thank you for attention! Link to Github

Skoltech
Skolkovo Institute of Science and Technology